Seminar der Graphischen Datenverarbeitung

Sommersemester 1998

Thema 11:

Video Rewrite

Arbeit von Christoph Bregler, Michelle Covell, Malcom Slaney,
Interval Research Corporation

Vortrag von Danilo Biella

Zusammenfassung

1. Einleitung

Video Rewrite ist ein Tool, welches ermöglicht, Filmmaterial so zu verändern, dass sich die Lippen eines Darstellers synchron zu einer zweiten Tonspur bewegen. Dies ist von Nutzen, wenn man einen Film vertont oder wenn man tote Personen etwas sagen lassen will, was sie niemals ausgesprochen haben. Ein Beispiel dafür ist der Auftritt von J. F. Kennedy im Film Forrest Gump. Allerdings wurden diese Bilder von Hand editiert. Video Rewrite ist das erste Gesichts-Animations-System, welches alles automatisch macht, mit Ausnahme von ein paar wenigen Dingen.
Der Mensch ist sehr empfindlich, was die Bild-Ton-Synchronisation anbelangt. Bewohner von nicht englischsprachigen Gebieten merken dies des öfteren. Wer sich in den Kinos deutsche Übersetzungen von amerikanischen Filmen ansieht, muss nicht nur die meist schlechte Umsetzung der Dialoge, sondern auch die unsynchronisierten Lippen ertragen.
Video Rewrite lernt von einem Trainingsset an Filmsequenzen, wie eine Person den Mund zu den einzelnen Tönen (Phoneme) bewegt. Später fügt es die einzelnen Teile so wieder zusammen, dass sie in die gewünschte Tonspur passen. Der Vorteil gegenüber mathematischen Modellen vom Kopf ist, dass individuelle Verformungen des Mundes mitgelernt werden.

2. Analyse der Trainingsdaten

Am Anfang stehen Filmsequenzen in denen der Darsteller sprechend vorkommt. Aus diesen Daten wird ein Video Modell erstellt, indem alle möglichen Artikulationen gespeichert sind. Dazu ist es notwendig das vorliegende Material zu markieren. Dies wird einerseits durch Bildanalyse, andererseits durch Tonanalyse bewerkstelligt. Als Ziel dieser Analysen steht ein Video Modell, welches eine Datenbank mit verschiedenen Triphonen darstellt.
Es ist notwendig, nicht mit einzelnen phonetischen Einheiten (Phoneme) zu arbeiten, sondern mit Triphonen (Tripletts von Phonemen), da das selbe Phonem je nach Kontext (vorheriges und folgendes Phonem) ein anderes Bild erzeugt. Zum Beispiel bewegt sich der Mund beim „b“ von „Bébé“ anders als bei „Bobo“. Würde man immer ein und dasselbe „b“ benutzen, sähe das Endresultat unnatürlich aus.

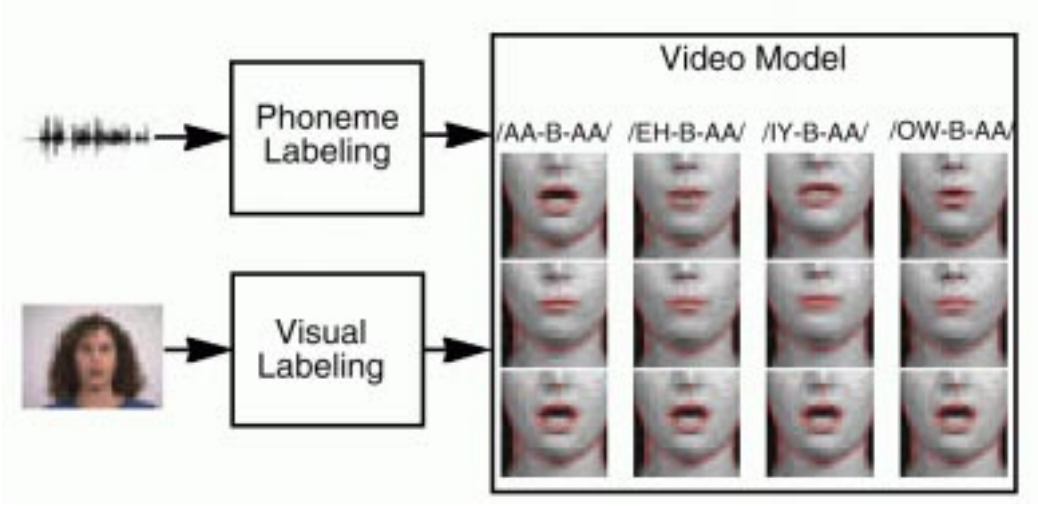

Fig. 1: Analyse der Trainingsdaten

2.1. Markierung durch Bildanalyse

Das Gesicht bewegt sich normalerweise rege innerhalb des Bildes, daher müssen wir immer herausfinden, wo der Mund ist und wie sich die Lippen bewegen. Randpunkte manuell zu markieren

ist ziemlich fehleranfällig und langwierig. Video Rewrite benutzt daher Computer-Vision Techniken um die Mundform und den Kiefer zu markieren.
Die Breite des Mundes beträgt nur ca. 40 Pixels. Konventionelle Contourtracking Programme versagen beim inneren Rand der Lippen. Hier wird darum der Eigenpoints Algorithmus bemüht, der mit wenig menschlichem Dazutun gute Resultate liefert. Eigenpoints lernt wie markierte Punkte sich im Bildverlauf verschieben und benutzt dieses Modell um neues Filmmaterial zu markieren. In diesem Beispiel wurden lediglich 26 Bilder von Hand markiert (0,2%).

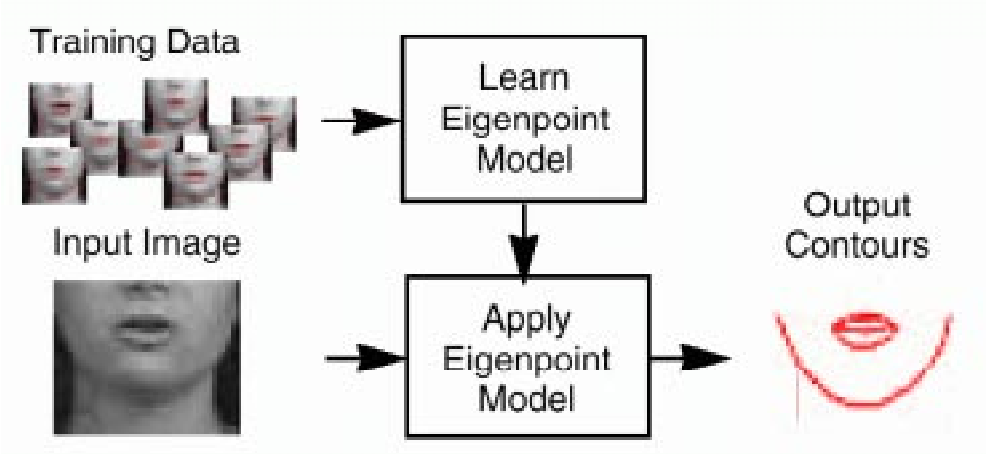

Fig.2: Funktionsweise von Eigenpoints

Im Fall von Video Rewrite werden 54 solcher „Eigenpunkte“ definiert. 20 für das Kinn und 34 für den Mund (20 für den äusseren Rand der Lippen, 12 Innen und je 1 für obere und untere Zähne).
Leider geht dieser Algorithmus nur von translatorischen Veränderungen in der 2D-Ebene aus. Das bedeutet, dass alle Bilder so gedreht werden müssen, dass der Mund in eine Referenzebene zu liegen kommt (und selber als 2D-Objekt angesehen werden muss).
Das Bild wird also affin abgebildet, so dass das Gesicht so gut wie möglich (least squares) in eine vordefinierte Maske passt.

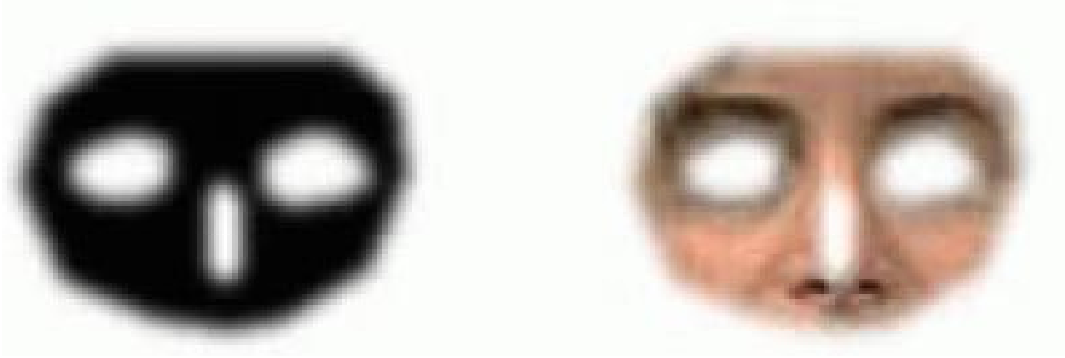

Fig.3: Maske für die globale Transformation und dazugehörende Gesichtspartie

2.2. Markierung durch Tonanalyse

Das ganze Trainingsset wird zu erst einmal in eine Phonem-Sequenz segmentiert. Da es sich aber hier um die Generierung einer Videospur handelt, sind nicht die Phoneme, sondern die Visime (Ein Visim ist der Teil des Films, der zum Phonem synchron ist) wichtig. Visim-Marken sind zeitlich an denselben Stellen zu finden wie die Phonem-Marken. Sie können aber gröberen Klassen zugeordnet werden. Rein visuell ist ein „P“ kaum von einem „B“, ja nicht einmal von einem „M“ zu unterscheiden (Man vergewissere sich vor dem Spiegel). Diese Phoneme werden daher auch in der gleichen Klasse zu finden sein. Es werden 26 Visimklassen definiert.

10 Klassen für Konsonanten:

(1) /CH/, /JH/, /SH/, /ZH/	(2) /K/, /G/, /N/, /L/
(3) /T/, /D/, /S/, /Z/	(4) /P/, /B/, /M/
(5) /F/, /V/	(6) /TH/, /DH/
(7) /W/, /R/	(8) /HH/
(9) /Y/	(10) /NG/

15 Vokalklassen:

(1) /EH/	(6) /AO/	(11) /OY/
(2) /EY/	(7) /AW/	(12) /IY/
(3) /ER/	(8) /AY/	(13) /IH/
(4) /UH/	(9) /UW/	(14) /AE/
(5) /AA/	(10) /OW/	(15) /AH/

und eine Klasse für Schweigen: /SIL/ (für Silence)

Aus diesen Marken macht Video Rewrite Triphon-Marken. Die Anzahl Marken wird durch diese Operation nicht vermindert. Die Triphone überlappen, d.h. zu jedem Phonem gibt es auch ein Triphon (am Rand sogar 1 mehr). Das englische Wort teapot sieht in Phonemmarken so aus:
/T/ /IY/ /P/ /AA/ /T/
Als Triphon-Sequenz dann so:
/SIL-SIL-T/ /SIL-T-IY/ /T-IY-P/ /IY-P-AA/ /P-AA-T/ /AA-T-SIL/ /T-SIL-SIL/.

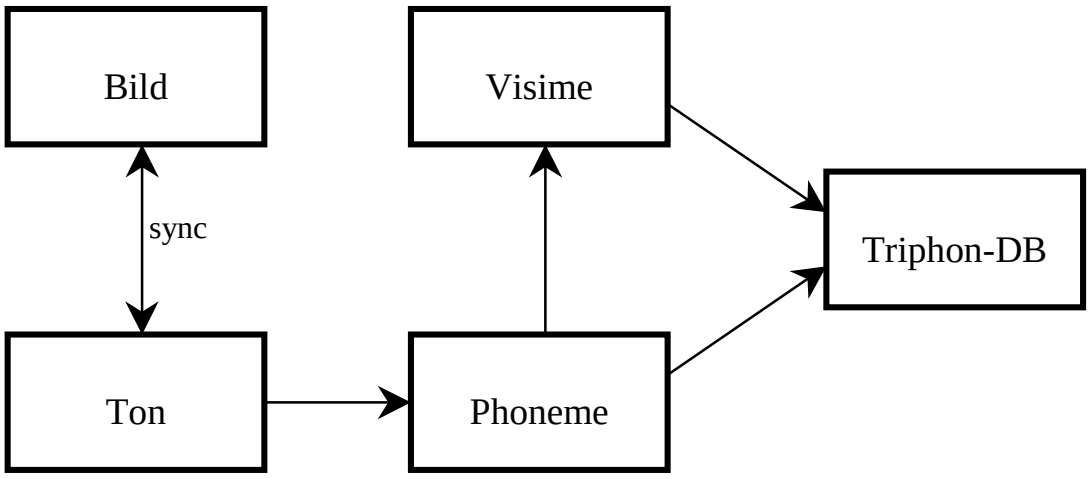

Fig.4: Tonanalyse

Die Markierung des Filmmaterials erfolgt mittels eines Hidden-Markov-Models.

3. Synthese der Triphone zu einer Videosequenz

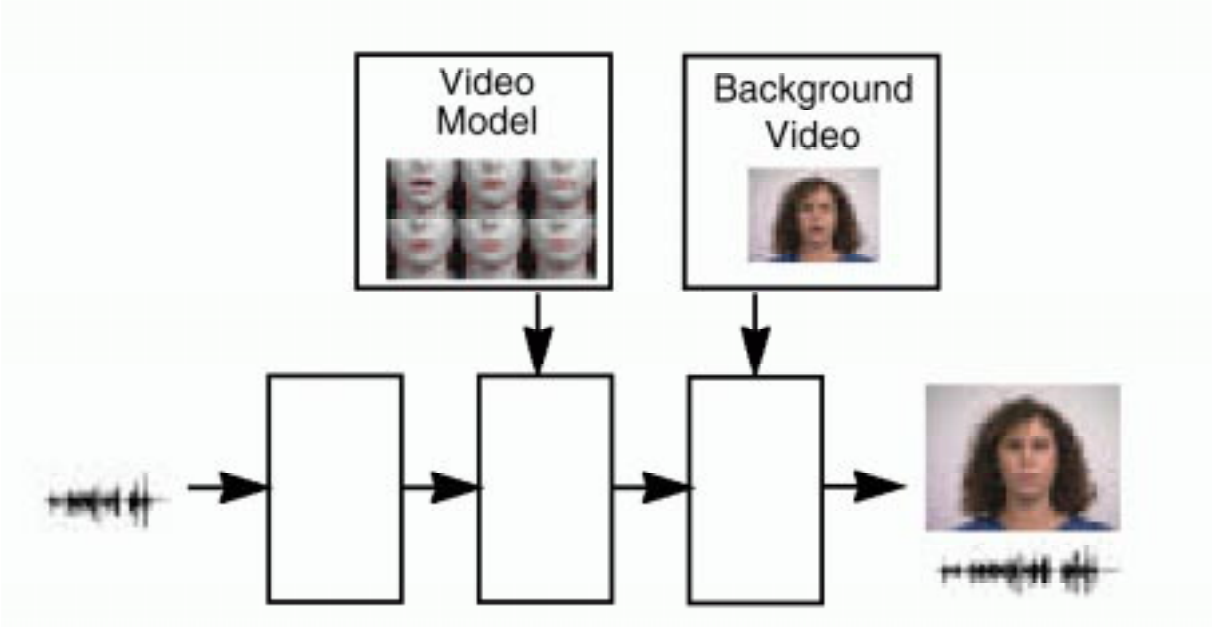

Fig.5: Synthese einer neuer Videosequenz

Video Rewrite synthetisiert das neue lip-sync-video indem die neue Tonspur wiederum Phonem-markiert wird und aus dem Video Modell die dazu passenden Stücke eingesetzt werden. Diese „Mundfilme“ werden dann an das Gesicht im Originalfilm ‚angenäht‘. Dieser Backgroundfilm bestimmt die Position, den Winkel und die Bewegung der Mundpartie. Der Rest des Gesichts wird belassen; Gestiken wie Blinzeln oder das Heben der Augenbrauen werden nicht beeinflusst. Die Mundbewegung und die Geschwindigkeit der Sprache werden hingegen von der Dubbing-Tonspur bestimmt. Die Triphone zeigen nur die Bewegungen, die direkt mit der Artikulation zu tun haben. Dies betrifft Mund, Kinn und Teile der Backen (Grübchen).

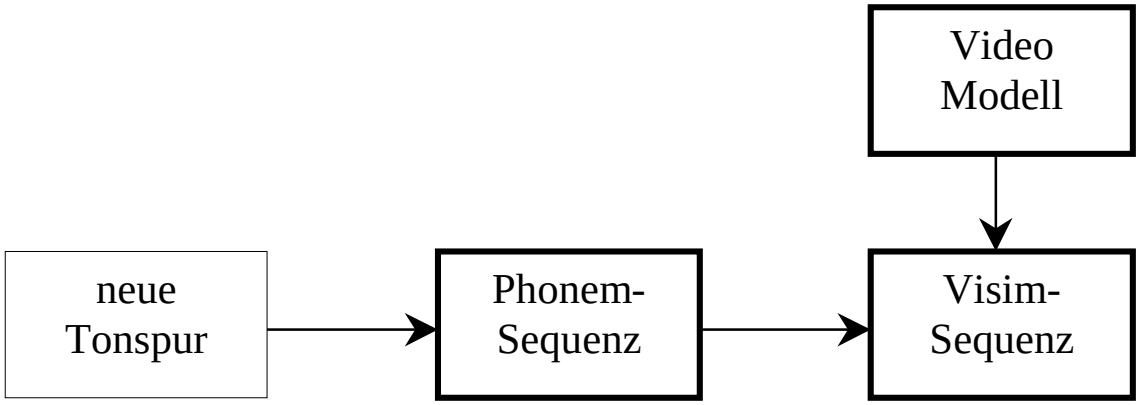

Fig.6: Auswahl der Videosequenz

3.1. Auswahl der richtigen Triphone

Als erstes wird die neue Tonspur markiert. Bei einem komplett ausgestatteten Video Modell könnte man die dazugehörigen Triphone auswählen und aneinanderreihen. Dies ist aber nicht der Fall. Es sind nicht alle möglichen Triphone im Trainingsset vorhanden. Trifft man in der Tonspur auf ein nicht erfasstes Triphon (im vorliegenden Fall kam dies zu 31% vor), so wird das nächstbeste ausgewählt. Zu jedem Triphon in der Datenbank wird der Fehler zum aktuellen errechnet. Dasjenige Triphon das diesen Wert minimiert, wird dann ausgewählt. Dieser Fehler wird durch zwei Parameter bestimmt: D_p und D_s. D_p ist die Phonem-Kontext-Distanz und D_s ist die Lippenform-Distanz von überlappenden Videobildern. Damit lautet die Fehlerformel:

$$\text{Error} = \alpha D_p + (1-\alpha) D_s,$$

wobei das Gewicht α die zwei Parameter gegeneinander aufwiegt. D_p stellt die gewichtete Summe der Phonem-Distanz zwischen Video Modell und Tonspur innerhalb des Triphons dar. Sind zwei Phoneme in der gleichen Phonemklasse (z.B. es sind beides /B/), dann ist die Distanz 0, sind aber beide in verschiedenen Visimklassen, so ist die Distanz 1. Für Phoneme in derselben Visimklasse werden Werte zwischen 0 und 1 angenommen. Das Mittlere Phonem hat das grösste Gewicht, gegen aussen nehmen die Gewichte ab. Es wird bei dieser Distanzberechnung nicht nur das aktuelle Triphon berücksichtigt, sondern auch ein Phonem vorher und nachher (Pentaphon). Da im Video Modell nur Triphone gespeichert sind, muss das aktuelle Triphon nicht nur mit allen Triphonen, sondern mit allen Trippeln verglichen werden. Der Rechenaufwand müsste um einige Grössenordnungen steigen, davon steht aber nichts im Paper.
Das D_s misst, wie gut die überlappenden Teile der Triphone zueinander passen. Dazu wird ein 4-dimensionaler Ortsvektor postuliert mit folgenden Einträgen: Äussere Lippenhöhe, äussere Lippenbreite, innere Lippenhöhe, innere Lippenbreite. Nun wird in jedem Bild innerhalb des überlappten Phonems die euklidische Distanz der Ortsvektoren der aktuellen Triphone berechnet. Diese Distanzen werden addiert. Um diese Summe zu minimieren werden die Triphone aussedem innerhalb von Schranken zeitlich gegeneinander verschoben. Das Minimum ist das D_s der jeweiligen Konfiguration.
Der Grund, warum man nicht einfach nur mit D_p rechnet ist folgender: Werden zum Vergleich nur die Triphonmarken herangezogen, so hat man nicht alle Details berücksichtigt. Zum Beispiel sagen die Marken noch nichts darüber aus, wie lange bei einem /P/ die Lippen zusammengepresst sind. Ausserdem werden die Marken von einem HMM gesetzt, dieser gibt keine 100% Garantie über die Korrektheit der Marken.
Ist die richtige Reihenfolge bestimmt, werden die Triphone mit Cross-Fading aneinandergereiht

3.2. Einpassen der Videosequenz

3.2.1. Innere zeitliche Anpassung

Die ausgewählte Sequenz muss jetzt in sich selbst abgestimmt werden. Durch das Minimieren von D_s wird die richtige Position der Marken zueinander und die optimale Dehnung der Phoneme bestimmt.

3.2.2. Zeitliche Anpassung an die Tonspur

Fig.7: Synchronisation mit der Tonspur

Die Triphonsequenz ist nun in sich selbst konsistent. Jetzt muss sie noch an die Tonspur angepasst werden. Der Startpunkt des jeweils mittleren Phonems im Triphon muss an die Stelle zu liegen kommen, wo sich die entsprechende Marke auf der neuen Tonspur befindet. Dazu werden die einzelnen Phoneme gestreckt oder gestaucht.

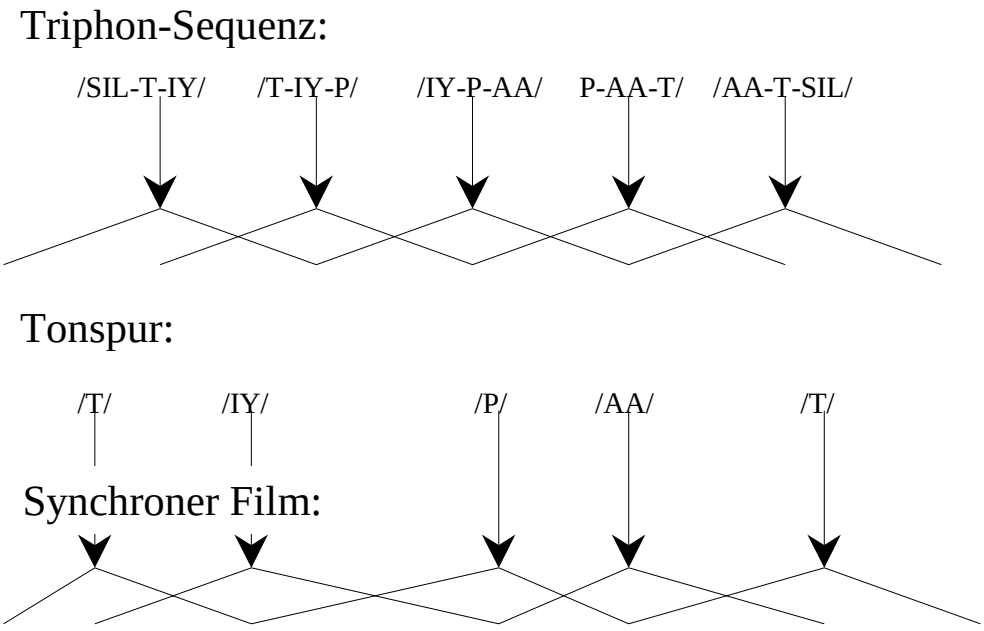

Fig.8: Skalierung der Triphon-Längen

3.2.3. Einbinden des Mundfilms in einen Hintergrund

Sowohl das Gesicht des Hintergrundfilms, als auch der Mund im neuen Teil bewegen sich, es ist notwendig diese richtig aneinander zu ‚heften', andernfalls würde der Mund im Endprodukt im Gesicht ‚herumschwimmen'. Die Maske aus Fig.3 wird nochmals benutzt um die globale Transformation durchzuführen. Eine zweite Maske (Fig.9) bestimmt dann, welcher Teil des Films aus dem Original und welcher Teil aus dem Mundfilm stammt.

Fig.9: Facial fadig mask

Weiter braucht es lokale Deformationen, um die Kieferlinie nicht zu zerstören. Der Unterkiefer im Original bewegt sich nicht so wie der der Mundpartie. Die Extreme (bei den Ohren und unter dem Mund) werden nicht verändert. In der Zwischenzone werden die Bildsignale gewichtet gemischt. Das Anheften wird durch die Mischung der zwei aktuellen Triphone mit den Hintergrundbild bewerkstelligt. Auch hier werden die Anteile individuell gewichtet.
Die berechneten Eigenpunkte werden als Kontrollpunkte für das Morphing benutzt. In jedem Endbild müssen 4 Bilder gewarped werden: die beiden Triphon-Bilder, die Maske und das Hintergrundbild.

4. Resultate

4.1. Methoden

Es wurden 8 Minuten Video aufgenommen, worin die Darstellerin insgesamt 109 Sätze aus Märchen zitiert. Zum Teil wurden Sequenzen mit ruhigem, zum Teil mit bewegtem Kopf aufgenommen. Das Video Modell wurde ausschliesslich mit Sequenzen trainiert, welche ruhige Bilder enthalten. Es enthielt dann 1‘700 Triphone. Um nicht zu optimistische Resultate zu erhalten, wurden als zweite Tonspur keine Sätze verwendet, die Ähnlichkeiten mit denen im Trainingsset aufwiesen (Zwei Sätze sind ähnlich, wenn zwei Wörter in beiden Sätzen hintereinander vorkommen). Video Rewrite würde in solchen Anwendungen bessere Resultate erzielen.

4.2. Beobachtungen

4.2.1. Lippensynchronisation

Bei der Lippensynchronisation wurden einige Timingprobleme beobachtet. Vor allem bei plosiven Lauten und bei Stops.

4.2.2. Triphon-Synchronisation

Bei schlechtem überlappen der Triphone kann es vorkommen, dass Flattern auftritt (Unstetigkeit). Beim Testen auf ruhigen Bildern wurden keine solchen Artefakte bemerkt.

4.2.3. Natürliche Artikulation

Unnatürliche Artikulierung kann auftreten, da nicht alle Triphone verfügbar sind, die auf Tonspur vorkommen. Beim Beispiel waren 31% nicht im Video Modell. Trotzdem wurden keine solchen Fehler beobachtet. Auch hier wurde auf ruhigen Originalbildern operiert.

4.2.4. Sichtbarkeit der Maske

Ohne illumination correction sieht man Artefakte um die Maske. Die Haut bewegt sich nicht richtig und man sieht Texturfehler. Bei Benutzung von illumination correction fallen diese aber wieder weg.

4.2.5. Hintergrund Warping

In manchen Teilen wurden Artefakte in der Gegend des äusseren Teils des Kiefers gefunden.

4.2.6. Räumlicher Zusammenhang

Der Mund ‚schwimmt‘ nicht, die Zähne sind fest am Kiefer angewachsen.

4.2.7. Gesamtqualität

Der lip-sync wird als excellent beurteilt.

5. Andere Experimente

5.1 Reduktion des Video Modells

Bei Reduzierung der Daten im Video Modell verschlechtert sich die Qualität des Resultats. Am Anfang, bei 1‘700 Triphonen (es existieren ca. 19‘000) werden nur 31% der Triphone in der Tonspur nicht in der Datenbank gefunden. Wird die Grösse des Video Modells halbiert, schlagen 46% der Suchen fehl. Bei weiteren Halbierungen betragen die Fehlschläge 58% und 74%.

5.2. Reanimation von historischem Filmmaterial

Video Rewrite wurde mit Kennedy’s Rede zur Kubakrise trainiert. Es wurden 2 Minuten Film (1157 Triphone) digitalisiert. 45 Sekunden sind Nahaufnahmen, der Rest von ein wenig weiter weg. Dieses Video Modell wurde benutzt, um Kennedy „I never met Forrest Gump“ oder „read my lips“ sagen zu lassen.
Die Qualität des Resultats ist nicht so gut wie im vorherigen Experiment. Dies hat zwei Gründe. Zum Ersten stehen hier weniger Triphone zur Verfügung und zum Zweiten bewegt sich der Kopf sehr stark in den Trainingsdaten. Die Lippenform wird verfälscht durch die affine Abbildung.

6. Verbesserungen, Ausblick

Es sind viele Erweiterungen denkbar. Die Phonem-Markierung berücksichtigt nur die akustischen Signale. Eine Markierung die die Lippenform einbezieht, wäre genauer.
Die akustischen Signale könnten benutzt werden, um die Triphone auszuwählen. Die Information in den Marken ist nicht stark genug, um Gesichtszüge zu beschreiben (Freude, Trauer ...).
Man könnte die Tonspur der Triphone gebrauchen, um den Film an die Zieltonspur anzugleichen, anstatt nur mit den Markierungen zu arbeiten.
Das Einbinden der Animation von anderen Gesichtsteilen (z.B. Augenbrauen) würde die Emotionen der Aussagen realistischer zum Ausdruck bringen. Man könnte nicht nur die Aussage, sondern auch die Stimmung des Darstellers verändern.